Duplication and divergence: The evolution of nematode globins

Journal of Nematology 41(1):35–51. 2009.

Duplication and Divergence: The Evolution of Nematode Globins

P. W. Hunt,[1,2*] J. McNally,[2] W. Barris,[3] M. L. Blaxter[1]

Abstract: In common with many other groups, nematodes express globins with unknown functions. Nematode globin-like genes can be divided into class 1 globins, similar to vertebrate myoglobins, and a wide range of additional classes. Here we show that class 1 nematode globins possess a huge amount of diversity in gene sequence and structure. There is evidence for multiple events of gene duplication, intron insertion and loss between species, and for allelic variation effecting both synonymous and non-synonymous sites within species. We have also examined gene expression patterns in class I globins from a variety of species. The results show variation in the degree of gene expression, but the tissue specificity and temporal specificity of expression may be more conserved in the phylum. Because the structure-function relationships for the binding and transport of oxygen by globins are well understood, the consequences of genetic variation causing amino acid changes are explored. The gene family shows great promise for discovering unique insights into both structure-function relationships of globins and their physiologial roles.

Key words: Genomics, Globin, Evolution, Genetic variation, Gene expression.

Globins are found ubiquitously in Archaea, Eubacteria and Eukaryota (Vinogradov et al. 2006). Globin genes have been identified in most fully sequenced archaeal, eubacterial and eukaryote genomes. However there are a few examples where globin genes have not been found, including some pathogenic bacteria. In the Metazoa, many members of the phylum Nematoda possess characterised globin genes, demonstrably similar to the well-known myoglobins of vertebrates (Blaxter 1993). Some of these genes have a signal peptide-encoding sequence, and the protein products have been demonstrated to be extracellularly located, while others are intracellular. These genes are grouped as class I nematode globins (Hoogewijs *et al.* 2007). Genome sequencing of *Caenorhabditis elegans* and close relatives has revealed the existence of a second class of globin-like genes that have just-detectable similarity to myoglobins, and specific roles in neural tissues (Hoogewijs *et al.* 2007). The roles of class I globins appear to lie mainly in oxygen binding, though whether this reflects a role as a transport protein or as some other system (such as enzymatic catalysis or sequestration) remains unclear (Vinogradov *et al.* 2008). One route to understanding and unravelling the physiology and evolution of globins is to survey their presence and variability across a wide range of taxa exhibiting different lifestyles, and to use this associative data to focus hypotheses of function.

Nematode globins have been the subject of previous analyses for three main reasons. The first is their abundance in adult nematode tissues (and in some secretions): many strongyle species are bright red (Blaxter et al. 1994a), and the mermithid *Mermis nigrescens* has a red, globin-containing 'eye spot' (Burr *et al.* 2000). The second is the discovery that some nematode class I globins have very different oxygen binding properties to their vertebrate homologues, showing, for example, thousand-fold increased affinities for oxygen. These properties made nematode globins a target for studies of general processes of haemoglobin structure-function relationships (Yang *et al.* 2005). Thirdly, globins were also identified as targets of host immunity in parasitic nematodes, a finding that may relate to both their abundance and presence in secretions (Frenkel *et al.* 1992; Vercauteren *et al.* 2003). The genomic structure of nematode class I globins also elicited interest. Vertebrate globins have a conserved three-exon structure, where the positions of the introns fit well with subdomains of the globin fold defined from structural studies. These intron positions have been used as strong evidence for an introns-early, genes-in-pieces model of protein evolution. Nematode globin genes were surprising because they had central introns that were clearly non-homologous to the two introns of vertebrate globins (Sherman *et al.* 1992; Moens *et al.*1992). The identification of two distinct central intron positions (in *C. elegans* and *Ascaris suum* globins) lent weight to an introns-late, insertional origin.

Most nematode class I globins isolated previously have been identified by directed searches. There is now a wealth of genomic and transcriptomic data for nematodes, and so we have used the resource of nematode expressed sequence tag (EST) projects and genome projects to identify many additional class I globin genes in a wide variety of parasitic and non-parasitic nematodes. In addition, we survey globin allelic diversity within one species (the strongyle *Haemonchus contortus*) to probe patterns of within-species diversification in class I globins. We also use EST and quantitative PCR analyses to examine tissue- and stage-specific expression patterns of class I globins in many species across the phylum.

Materials and Methods

Nomenclature: We name each globin gene identified with-the three letter gene name *glb* preceeded by

Received for publication December 18, 2008.
[1]Institute of Evolutionary Biology, University of Edinburgh, King's Buildings, Edinburgh EH9 3JT, United Kingdom.
[2]CSIRO, Locked Mail Bag 1, Armidale, NSW 2350, Australia.
[3]CSIRO, 306 Carmody Road, St Lucia QLD 4067, Australia.
The authors gratefully acknowledge the gift of biological materials from K. Tetteh and E. Riga, and the sharing of sequence information from L. Liu, and preliminary data by D. Hoogewijs, S. McClure and D. Emery, technical assistance from J. Kenny and funding support from the Leverhulme Trust and CSIRO. The genomic data for *H contortus* were generated by Matt Berrimann and colleagues of the Wellcome Trust Sanger Institute, Cambridge UK. Supplementary material for this article is available at the website: http://www.nematodes.org/downloads_area/supplementary_information/Hunt_2009/peter.hunt@csiro.au
This paper was edited by Paula Agudelo.

a three letter abbreviation of the species name (e.g. *Tca-glbm* for *Toxocara canis* myoglobin), following precedence (Blaxter 1997). We also refer to globin cDNAs with this italicised notation. For proteins, we follow the SwissProt format: Roman uppercase followed by a five letter species abbreviation (e.g. GLBM_TOXCA). Where genes and proteins can be classified into one of the known subgroups of class I nematode globins they are further designated: *glbm*/GLBM for body wall or intracellular globin, *glbc*/GLBC for cuticular, secreted globin, *glbp*/GLBP for pseudocoelomic globin (and, as some ascaridid nematode GLBP proteins consist of two distinct globin-like domains, these are further distinguished as GLBA and GLBB), and *glbe*/GLBE for eye globin. For globins of uncertain subgroup affinity, we follow the name with the cluster or contig number (e.g. *Tca-glb*_cn00408; see below for definition of clustering and contigs). When discussing gene structure and protein residues, we have used the alpha-numeric coding system for referring to helices and interhelix residues. We have followed the alignment of GLBA_ASCSU (*Ascaris* D1) and GLBM_PHYCA (*Physeter catodon* [sperm whale] myoglobin) based on crystal structures presented in Yang *et al.* 1995 for our alignment of amino acid sequence (Supplementary Table 2), and residues are named after the corresponding GLBM_PHYCA residues. There are a number of places in the alignment where there is no corresponding GLBM_PHYCA residue. Where these amino acid residues are mentioned, we name them after the preceeding GLBM_PHYCA residue followed by a letter as in the ef interhelix in Figure 2 and supplementary Table 2.

Directed isolation of new class I nematode globin genes Syngamus trachea *globin genes: S. trachea* adult pairs (a gift of Dr. E. Riga, Aberystwyth [now at Washington State University]) were obtained from fledgling crows and frozen at -80°C. Genomic DNA, RNA and protein were prepared simultaneously from four pairs using TRISOLV reagent (Biotecx Laboratories, Inc.). Total RNA was reverse transcribed into first strand cDNA and PCR performed with degenerate primers SG1 and SG3 designed from an alignment of previously isolated strongylid globins (Blaxter et al. 1994a; Vanfleteren et al. 1994) and, oligo d(T). After sequencing the amplicons obtained, additional reverse primers were designed and used with the nematode spliced leader, SL1 (5′-GGTTTAATTATCCCAAGTTTGAG-3′) to obtain the 5' part of the cDNA. Full length cDNA was subsequently isolated with specific primers. Analysis of the sequences revealed two distinct genes had been cloned. The two genes cloned (*Str-glbm* and *Str-glbc*) were identified as a myoglobin and a cuticle globin respectively. Genomic DNA corresponding to these two loci was isolated by PCR and cloned before being sequenced on both strands.

Toxocara canis *globin genes:* Whole adult *T. canis* (a gift of K. Tetteh, Edinburgh) were ground in liquid nitrogen and RNA extracted using the Ultraspec RNA isolation system (Biotecx Laboratories, Inc.). Oligo-dT primed cDNA was generated using the GeneAmp™ kit (Perkin Elmer). Body wall globin cDNAs (*Tca-glbm*) were isolated by amplification of first strand cDNA using an upstream primer TcG-S (5'-AATGGCGAC GGCATGCTTG-3') based on partial cDNA sequence (kindly provided by Dr L. Liu, Harvard) and a GC-anchored oligo d(T) primer DGDT (5'-GCGCGGAT CCGCTTTTTTTTTTTTTTTTTT-3'). This PCR product was cloned and sequenced. The *T. canis* pseudocoelomic globin cDNA (*Tca-glbp*) was amplified using degenerate primers (APG-dF 5'–TAYAARCAYATGTT YGARMAYTAY CC-3' and APG-dR 5'-GGRTARTKYT CRAACATRTGYTTRTA-3'; derived from an alignment of the four previously published ascaridid pseudocoelomic globin domains (Dixon et al. 1991, Sherman et al. 1992)), in combination with DGDT and SL1 primers respectively. PCR products (approximately 200 bp for APG-dR - SL1 and 500 bp for APG-dF - DGDT) were cloned and sequenced. To confirm the size and sequence of the predicted cDNA a second RT-PCR experiment was conducted using a new 5' primer TcpG-S (5'-TGCGATCTTTTGCGTTGTTTG-3') and DGDT. This PCR product was also cloned and sequenced. Both *T. canis* globin transcripts are *trans*-spliced to SL1. Genomic DNA from individual *T canis* nematodes was amplified with primers derived from cDNA, and fragments cloned and sequenced. Introns were predicted with reference to the relevant cDNA sequence. Intron position in the proteins was derived through alignment of the encoded proteins with the *A. suum* pseudocoelomic globin domain A (GLBA_ASCSU), for which the three dimensional structure has been resolved (Yang et al. 1995).

Identification of class I nematode globin genes from EST data: Globin-like genes were identified in EST datasets clustered into putative genes using PartiGene in the NEMBASE3 database (Parkinson et al. 2004; Wasmuth *et al.* 2008) or, for *H. contortus* data, StackPACK in WormSIS - CSIRO. For some analyses, multiple clusters assembled algorithmically were collapsed together to form new consolidated clusters by eye. This was especially necessary for some *Teladorsagia circumcincta* clusters. A description of the EST clusters is given in Table 1a, and Supplementary Table 1 lists the GenBank accessions for ESTs comprising the clusters. From each cluster of ESTs a consensus sequence or contig was predicted. Table 1b lists other globin sequences used in the phylogenetic and protein structure analyses described below. Contigs from WormSIS and those which have been adjusted are indicated in Table 1a: these sequences can be provided by the authors on request. NEMBASE3 data are freely available at http://www.nematodes.org/nembase3.

Phylogenetic analysis: We assembled an alignment of inferred amino acid sequences from the EST clusters

TABLE 1. Nematode globin sequences used for clustering, phylogenetics and SNP identification

(A) – Nematode globins from EST clusters.

Species	Cluster[a]	Cluster origin	Number of ESTs[d]	SNPs identified[b]
Ancylostoma caninum	cn01487	NEMBASE	2	excluded
Ancylostoma caninum	cn03829	NEMBASE	3	excluded
Ascaris lumbricoides	cn00043	NEMBASE	41	1
Ascaris suum	cn00004[c]	NEMBASE	82	13
Ascaris suum	cn00019[c]	NEMBASE	5	excluded
Ascaris suum	cn00780	NEMBASE	7	2
Ascaris suum	cn17423	NEMBASE	2	excluded
Ancylostoma ceylanicum	orphan	NEMBASE	1	excluded
Ancylostoma ceylanicum	cn00272	NEMBASE	13	0
Ancylostoma ceylanicum	cn00544	NEMBASE	4	excluded
Ancylostoma ceylanicum	cn01816	NEMBASE	2	excluded
Caenorhabditis elegans	Clones of ZK637.13[c]	WormBase	18	0
Haemonchus contortus	orphans	NEMBASE	2	excluded
Haemonchus contortus	cn00868[e]	WormSIS	24	43
Haemonchus contortus	cn01319[e]	WormSIS	9	15
Haemonchus contortus	cn01320[e]	WormSIS	29	66
Haemonchus contortus	cn01356[e]	WormSIS	2	excluded
Haemonchus contortus	cn01377[e]	WormSIS	23	53
Haemonchus contortus	cn01583[e]	WormSIS	4	excluded
Haemonchus contortus	cn01747[e]	WormSIS	15	7
Haemonchus contortus	cn08501	NEMBASE	3	excluded
Heterodera glycines	orphan	NEMBASE	1	excluded
Heterodera glycines	cn06223	NEMBASE	5	1
Litomosoides sigmodontis	orphan	NEMBASE	1	excluded
Meloidogyne chitwoodi	cn00083	NEMBASE	2	excluded
Meloidogyne hapla	cn02564	NEMBASE	3	excluded
Meloidogyne incognita	cn00586	NEMBASE	5	0
Meloidogyne javanica	orphan	NEMBASE	1	excluded
Necator americanus	orphan	NEMBASE	1	excluded
Necator americanus	cn00041	NEMBASE	3	excluded
Necator americanus	cn00088	NEMBASE	2	excluded
Nippostrongylus brasiliensis	orphans	NEMBASE	1	excluded
Nippostrongylus brasiliensis	cn00124	NEMBASE	2	excluded
Nippostrongylus brasiliensis	cn00197[c]	NEMBASE	19	17
Nippostrongylus brasiliensis	cn00328	NEMBASE	7	1
Ostertagia ostertagi	orphan	NEMBASE	1	excluded
Ostertagia ostertagi	cn00190	NEMBASE	40	38
Ostertagia ostertagi	cn00214	NEMBASE	2	excluded
Ostertagia ostertagi	cn03092	NEMBASE	2	excluded
Ostertagia ostertagi	cn03348	NEMBASE	2	excluded
Onchocerca volvulus	cn00634	NEMBASE	11	2
Parastrongyloides trichosuri	orphan	NEMBASE	1	excluded
Strongyloides stercoralis	cn00231	NEMBASE	4	excluded
Strongyloides stercoralis	cn02226	NEMBASE	8	1
Toxocara canis	cn0537[c]	NEMBASE	8	1
Toxocara canis	cn0563[c]	NEMBASE	4	excluded
Teladorsagia circumcincta	orphan	NEMBASE	1	excluded
Teladorsagia circumcincta	cn00008[e]	NEMBASE	34	57
Teladorsagia circumcincta	cn00009[e]	NEMBASE	23	52
Teladorsagia circumcincta	cn00032	NEMBASE	4	excluded
Teladorsagia circumcincta	cn00084	NEMBASE	3	excluded
Teladorsagia circumcincta	cn00173	NEMBASE	8	1
Teladorsagia circumcincta	cn01113	NEMBASE	9	4
Trichuris muris	cn00180	NEMBASE	6	0
Trichuris muris	cn01615	NEMBASE	2	excluded
Trichinella spiralis	cn03467	NEMBASE	3	excluded
Trichinella spiralis	cn00408	NEMBASE	13	0
Trichuris vulpis	cn00077	NEMBASE	3	excluded
Trichuris vulpis	cn00688	NEMBASE	5	0
Xiphinema index	cn00721	NEMBASE	3	excluded
Zeldia punctata	orphan	NEMBASE	1	excluded

[a] Some ESTs were not clustered with any other sequence. These are called "orphans" in this column of the table.

[b] Where there were fewer than 5 ESTs in a cluster, these were not used to infer SNPs and are identified as "excluded" in this column of the table (see materials and methods).

[c] These clusters encode cDNA which is near identical to genes described in Table 1b.

[d] For GenBank accessions, see Supplementary Table 1.

[e] These clusters differed from those presented in NEMBASE, see materials and methods.

(B) – Other nematode globins.

Species	Gene Name*	GenBank Accession(s)	Corresponding cluster	Reference
Ascaris suum	*As-glbm*	U17337	cn00004	(Blaxter *et al.* 1994b)
Ascaris suum	*As-glbp* (*GLBA, GLBB*)†	L03351	cn00019	(Sherman *et al.* 1992)
Brugia malayi	*Bma-glb*	NW_001893041	N/A	(Hoogewijs *et al.* 2008)
Caenorhabditis briggsae	*Cbr-glb*	U48289,U48290, U48291	N/A	(Kloek *et al.* 1996)
Caenorhabditis elegans	*Ce-glb-1*	NM_066573	N/A	(Kloek *et al.* 1996)
Caenorhabditis remanei	*Cre-glb*	U48295, U48294, U48293, U48292	N/A	(Kloek *et al.* 1996)
Mermis nigrescens	*Mni-glbm*	AF138291, AF138296	N/A	(Burr *et al.* 2000)
Mermis nigrescens	*Mni-glbe*	AF138295, AF138294, AF138293, AF138292,	N/A	(Burr *et al.* 2000)
Nippostrongylus brasiliensis	*Nbr-glbm*	L20895, L25872, L25873	cn00095**	(Blaxter *et al.* 1994a)
Nippostrongylus brasiliensis	*Nbr-glbc*	L20896, L25874	cn00197**	(Blaxter *et al.* 1994a)
Ostertagia ostertagi	*Oci-glbc*	AJ427357	none	(Vercauteren *et al.* 2003)
Pseudoterranova decipiens	*Pde-glbp* (*GLBA, GLBB*)†	M63298	N/A	(Dixon *et al.* 1992)
Syngamus trachea	*Str-glbm*	AF370722	N/A	This report
Syngamus trachea	*Str-glbc*	AF370721	N/A	This report
Toxocara canis	*Tca-glbp*	AF370726	cn00537	This report
Toxocara canis	*Tca-glbm*	AF370724, AF370723	cn00563	This report
Trichostrongylus colubriformis	*Tco-glbc*	M63263	N/A	(Frenkel *et al.* 1992)

* - Name used in this report, ** - Not an exact match, † - Di-domain globins domains are separated for phylogenetic and structural analyses (Figures 1 and 2).

(see next section) and nematode globins from GenBank. The full alignment is available in Supplementary Table 2. Predicted amino acid sequences from the consensus of each cluster were aligned by eye. For phylogenetic reconstruction, globins represented by single ESTs (orphans in Table 1a) were not used, as the quality of the predicted amino acid sequence could not be affirmed. We did not utilise the genome-project predicted *Pristionchus pacificus* globin sequences (see http://www.pristionchus.org) as our predictions disagree with those derived computationally (See Supplementary Figures 4-7. Preliminary trees were generated using a neighbour joining (using default settings in PAUP*, Unix version (Swofford 2000)), and these showed that the sequences from genera *Trichinella*, *Trichuris, Xiphinema* and *Mermis* formed a separate group. This observation concurrs with traditional and molecular phylogenetic analyses, which place these taxa in the Dorylaimia, distinct from all other taxa studied, which are in the Chromadorea (Meldal *et al.* 2007). This was also observed in analyses using maximum parsimony and so this group was defined as the outgroup for further analyses. These subsequent analyses were undertaken using Bayesian, parsimony and neighbour joining methods. Bayesian analyses were carried out on the portion of the alignment corresponding to the globin domain of the *A. suum* myoglobin (i.e. excluding the secretory leader peptides and polar zipper extensions) in MrBayes 3.1.2 (Ronquist and Huelsenbeck 2003). A flat prior was assumed on the underpinning amino acid substitution model, and two chains of MCMC analysis were run for 2.5 million generations. Inspection of the chains, sampled every 100 generations, in Tracer (version 1.4.1; A. Rambaut; http://tree.bio.ed.ac.uk/software/tracer/) showed that stationarity had been achieved after ~50,000 generations, and the split frequency between the chains was less than 0.001. A consensus tree was derived from the last 2.4 million generations, and Bayesian posterior probability support estimated. The consensus tree presented in figure 1 was examined and annotated in FigTree (version 1.2.2; Andrew Rambaut; http://tree.bio.ed.ac.uk/software/figtree/). Maximum parsimony analyses were carried out on the reduced alignment (as for MrBayes) in PAUP* version 4.10b (Swofford) by full heuristic search, with reliability of the resulting tree estimated by performing 10,000 bootstrap resamplings. Neighbour joining analyses were carried out in PAUP* version 4.10b (Swofford) using the BioNJ method, with reliability of the resulting tree estimated by performing 10,000 bootstrap resamplings.

Identification and verification of single nucleotide polymorphisms (SNPs) in globin genes Prediction of SNPs from EST clusters: An analysis program (findsnps1.pl) was written using perl. The program interrogates alignments, ignoring sequence differences occurring in the beginning and ending of each sequence in the alignment in order to minimize the effects of low quality sequence. SNPs are predicted when one or more sequences have non-consensus base pairs at a position in the alignment. For the results reported in this paper, we report predicted SNPs only from those alignments where there were five or more ESTs, and also from only those positions in alignments where there were five or more sequences aligned. Because of the real risk of a single observation being due to sequencing errors, rather than true allelic polymorphism, the number of observed alternative bases had to be found in 2 or more sequences for the SNP to be considered. For all of our alignments (the largest cluster contains 73 sequences), this "two-or-more" rule approximates rejecting each SNP if the lower limit of the 99% confidence interval for the minor allele frequency drops below zero. We also excluded SNP polymorphisms from the data where

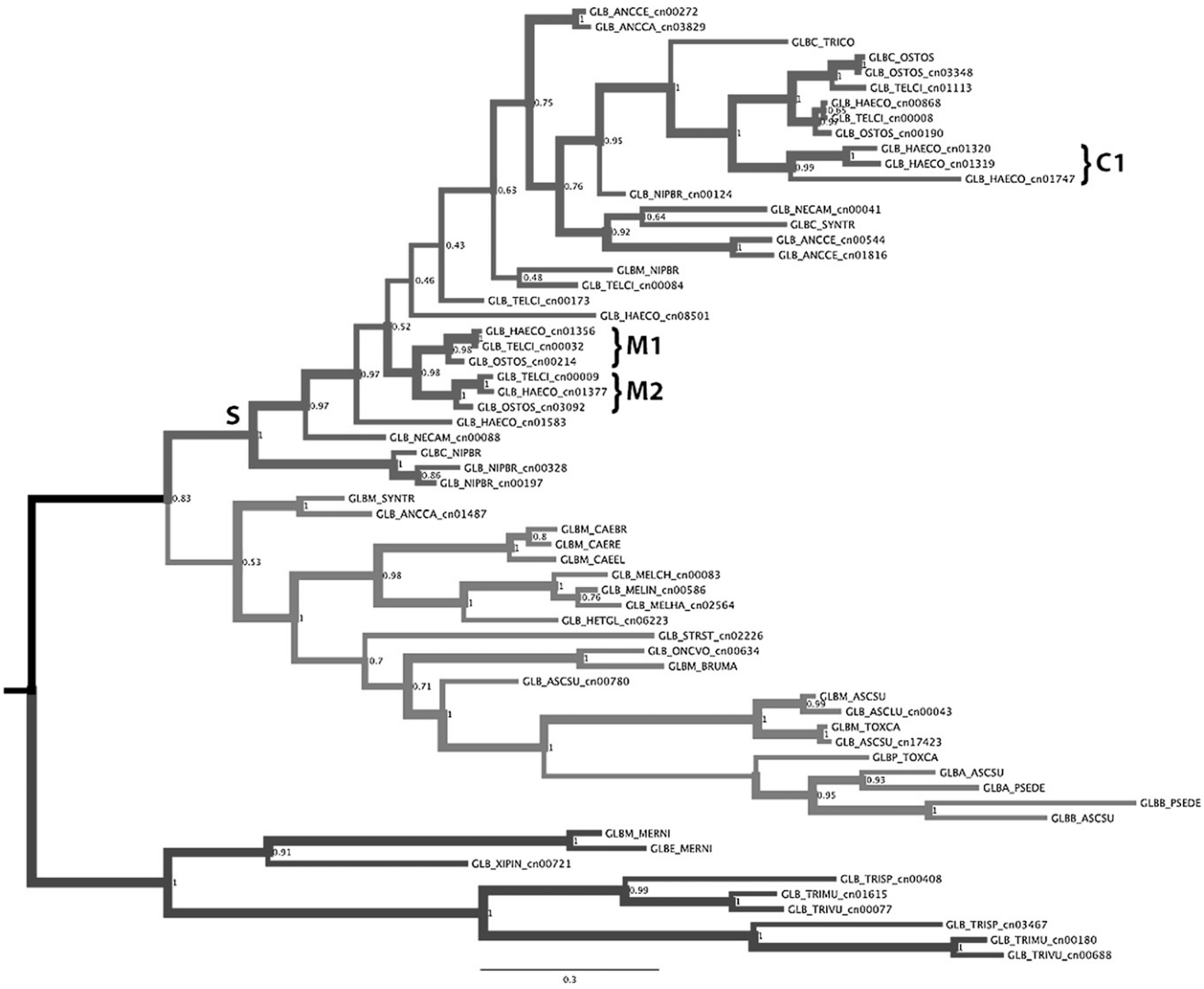


FIG. 1. Phylogenetic relationships between the EST cluster consensus sequences for 47 putative nematode globins and 15 nematode globins from GenBank. A tree of these globins is presented, based on Bayesian analysis of the amino acid sequence alignment. The numbers at the nodes represent the posterior probability support for the node. Branch lengths are proportional to evolutionary distance, and branch thickness is greater with increasing statistical support. Trees were also generated using maximum parsimony and minimum evolution methods and these are presented in Supplementary Figures 1 and 2 respectively. Gene duplications and a large amount of divergence combined with a sparse sampling throughout the phylum result in trees with poor resolution, though some groupings are observed with high bootstrap support. Clades discussed further in the text are indicated with letters and brackets. Group S is a group containing almost all Strongylid globins, group C1 contains three highly similar putative extracellular globins from *H. contortus*, and groups M1 and M2 show two clades of Strongylid globins where each contains one sequence each from *H. contortus*, *T. circumcincta* and *O. ostertagi*.

there was more than one minor allele and none of these were observed more than once. As a consequence, we have reported some SNPs to be bi-allelic when there has been only a single observation of third or fourth alternative alleles.

Confirmation of SNPs in two Haemonchus contortus *globin genes:* Analyses of *H. contortus* globin clusters showed that *Hco-glb*_cn01320 and *Hco-glb*_cn01319 were very similar, and that both had a large number of predicted SNPs. We attempted to amplify fragments of both these genes from three individual *H. contortus*. DNA from two males and the head region of one female *H. contortus* adults were extracted using Chelex resin (BioRad) as described previously (Hunt et al. 2008). The primers cn1320-F (5'-AAAGCTTCACTGCCGATGAC-3') and cn1320-R2 (5'-AGCATGAGGGAGTCCAAGAG-3') were used to amplify a segment of *Hco-glb*_cn01320 and the primers cn1319-F (5'-CCATCCAGAAAATCGCAAAT-3') and cn1319-E (5'-CCATTATACATGTGGAAGACCG AGGCTTGCGAGATG-3') were used for *Hco-glb*_cn01319. As part of the work it was determined that cn1320-F is not specific for *Hco-glb*_cn01320, and can be used to amplify a segment of *Hco-glb*_cn01319 using cn1319-E as a reverse primer. PCR products amplified using proof reading *Pfu* DNA polymerase (Promega) were subsequently cloned into the pGEM-T easy vector (Promega). Severalclones (Figure 4A) from each amplification were sequenced and aligned by eye with both the EST

cluster consensus and individual reads from the *H. contortus* genome sequencing project (http://www.sanger.ac.uk/Projects/H_contortus/). The alignment was analysed using PAUP* as described above. By aligning the whole genome shotgun (WGS) sequences obtained with the EST cluster alignment, three categories of SNPs were identified; those which were evident in both alignments, those in the EST alignment only and those in the WGS only.

Gene expression analyses Multiplex RT-PCR: T. canis tissues for RT-PCR experiments were obtained by dissecting live nematodes, and rapidly freezing the tissues in liquid nitrogen. Frozen tissue was either ground in liquid nitrogen (whole females, males and head sections) or homogenised in Ultraspec RNA reagent using microhomogenisers. RNA and resultant cDNA was then isolated as above. Multiplex RT-PCR for detecting globin transcripts in *T. canis* tissue was performed by co-amplifying either *Tca-glbm* or *Tca-glbp* with *T. canis* ribosomal protein L3 (*Tca-rpl-3*) (GenBank accession U17358) transcript as an internal control. Primers used were TcpG-S and TcpG-E1 (5'-GGTGGCTATGACT GCTTTCATGTTG-3') for *Tca-glbp* or TcG-S and TcG-E (5'-GAAATGGTCTAATGGGGT-3') for *Tca-glbm* with TcR-1F (5'-CGTTTATCGCATTCAAGGCTGG-3') and TcR-2R (5'-GCAATCCTCGCTAAGATGTTCAGC-3') for *Tca-rpl-3*. Products were analysed on agarose gels.

Expression levels of EST clusters: For genes identified in the EST datasets, the tissue and stage of origin of the libraries from which the ESTs were derived was ascertained, and these data used to estimate tissue- and stage-specificity of expression of each gene. Expression levels derived in this way are expressed as proportions of the total numbers of ESTs generated from the relevant libraries.

Results

Toxocara canis *globin genes:* The dog ascaridid *T. canis* lives as an adult in the dog gut, but infective larvae can cause human disease if eggs are ingested. The presence of globins in this nematode was not known, but the related *A. suum* has two well characterised isoforms. Two globin genes, *Tca-glbm* and *Tca-glbp*, were identified. The predicted protein GLBM_TOXCA has no secretory leader peptide and a high amino acid identity (84%) to *A. suum* myoglobin (GLBM_ASCSU). The central region (B6 to F9) of the predicted protein sequences differ by ten residues, five of which are highly conservative substitutions (Figure 2). This classifies GLBM_TOXCA as an intracellular, tissue or myoglobin type molecule. In contrast GLBP_TOXCA has a secretory leader sequence and is highly similar to the perienteric fluid hemoglobins of *A. suum* (Sherman et al. 1992) (GLBP ASCSU; GLBA domain 75% identity; GLBB domain 67% identity) and *Pseudoterranova decipiens* (Dixon et al. 1991) (GLBP PSEDE; GLBA domain 70% identity; GLBB domain 60% identity). GLBP_TOXCA however, has only a single globin domain and lacks the predicted polar zipper sequence (de Baere et al. 1992) possessed by GLBP_ASCSU and GLBP_PSEDE. GLBP_TOXCA is classified as a perienteric (pseudocoelomic) fluid type hemoglobin.

Syngamus trachea *globin genes:* The strongylid parasitic nematode *S. trachea* lives in the airways of avian hosts, a highly oxygenated environment, so globins from this species were cloned to enable comparison with other strongylids such as *H. contortus* which parasitise the micro-aerobic, lumenal surface of the gastrointestinal tract. PCR was performed on oligo-dT primed (primer DGDT) first strand cDNA using primers SG1 and DGDT, and SG3 and DGDT. Two products of differing size were amplified and cloned. Sequencing revealed that these corresponded to distinct products both with nucleotide sequence similarity with nematode globins. Oligonucleotides designed to the unique regions downstream from the predicted termination codon in each gene were then used in PCR with SL1 as the upstream primer and SL1-DGDT cDNA as target. Products of the expected size (~500 base pairs) were cloned and sequenced and the complete cDNA sequences of the two transcripts assembled.

Reverse transcription and SL1-DGDT amplification of *S. trachea* poly(A)+ RNA resulted in a smear of products from ~100 bp to >3 kb, with two more abundant bands discernable. As the globins are abundant protein products of the nematodes, we reasoned that the abundant transcripts might encode them. The larger of the abundant bands (~700 bp) was excised and blunt end cloned. Several clones were sequenced and proved to encode distinct proteins including a homologue of the *Nippostrongylus brasiliensis* Hsp20 gene (Tweedie et al. 1993) and a gene with a cysteine-rich repeat motif with homologues in *C. elegans* and *Brugia malayi* (Fuhrman et al. 1995; Blaxter 1996), as well as globin transcripts. Two globin cDNAs were identified. *Str-glbm* encodes a protein of 158 amino acids and *Str-glbc* encodes a protein of 170 amino acids, which has an N-terminal hydrophobic extension predicted to be a secretory signal peptide.

Nematode globins defined by clustering ESTs: Using an E-value cutoff of $1x10^{-6}$, we distinguished 85 clusters from NEMBASE3 (http://www.nematodes.org/nembase3/) (Wasmuth et al. 2008) that had significant matches to HMMPfam IPR000971 (Table 1). These genes are from 25 species, with the number of clusters for each taxon varying between 12 (*T. circumcincta*) and 1 (12 species). A cluster of *C. elegans* ESTs corresponding to the globin gene *Ce-glb-1* (ZK637.13) was also included for comparison. The total number of ESTs available for each species varies, and for those with low numbers it is possible that further globin sequences remain undefined.

Protein predictions derived from ascaridid (*A. suum*, *A. lumbricoides* and *T. canis*) clusters were aligned with

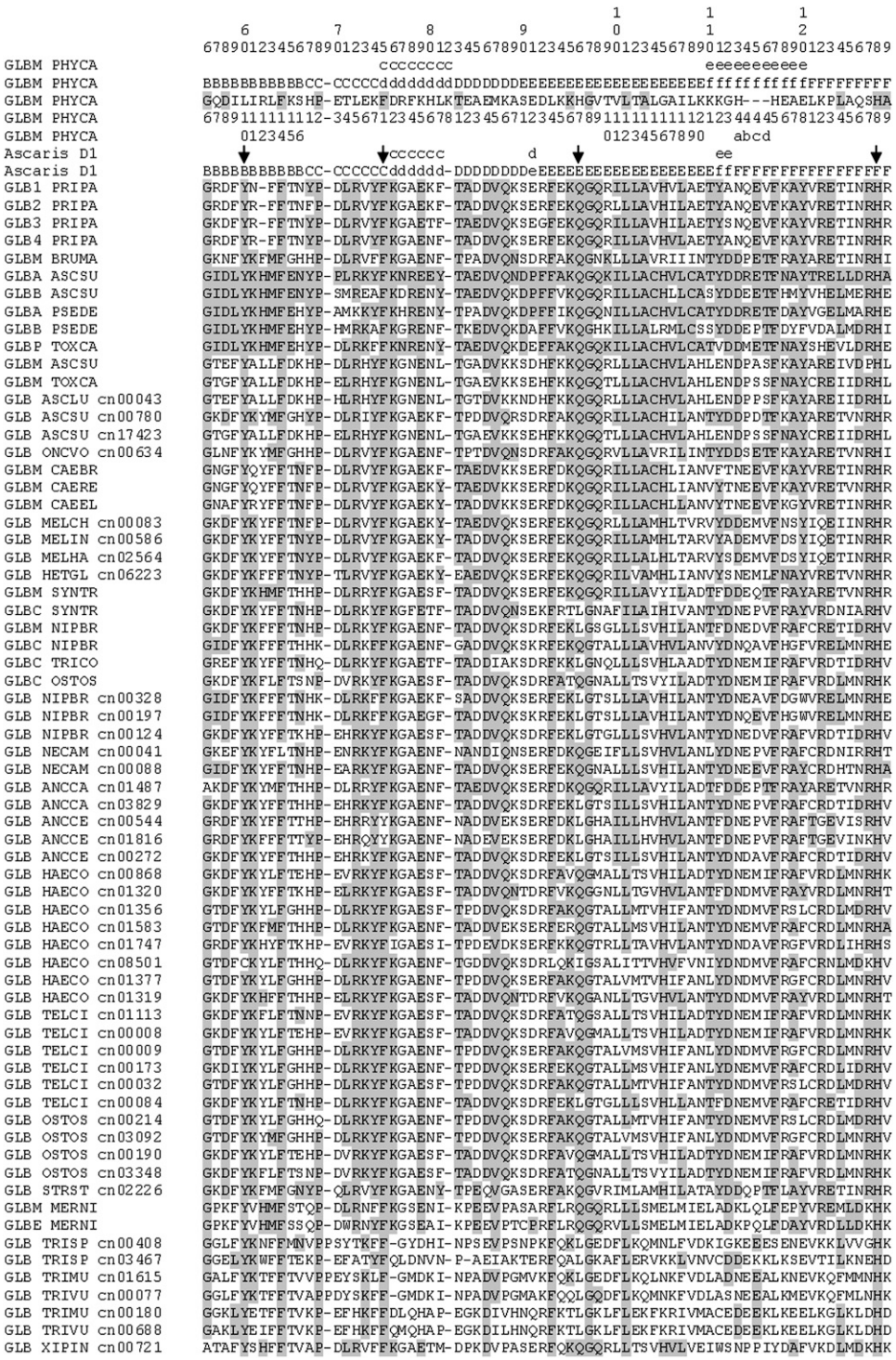


FIG. 2. Partial alignment of nematode globin sequences showing homology to GLBP_ASCSU domain A with identical residues shaded grey. The three dimensional structure of this gene "Ascaris D1" is illustrated directly above the alignment, and the three dimensional structure and sequence of *Physeter catodon* myoglobin GLBM_PHYCA is shown above D1. Helix numbering below the Helix designation letters follows the *P. catodon* structure. The numbering at the top of the figure indicates the portion of the alignment from which the figure is derived. The entire alignment is available as Supplementary Table 2. Arrows indicate the key ligand coordinating residues B10 and E7. The heme coordinating residue F8 and the ligand pocket gating residue cd1.

GLBM_TOXCA and GLBP_TOXCA. Tentatively, this EST clustering shows that *A. suum* has one additional extracellular globin quite distinct from the published sequence (*As-glb*_cn00780).

H. contortus globins were originally divided into 22 clusters, but re-analysis using an alternative clustering algorithm placed these into 8, based on more conservative criteria. For *H. contortus*, the globin EST cluster

consensus sequences were analysed using SignalP (Emanuelsson et al. 2007) to check for secretory leader peptides. The results indicate that *Hco-glb*_cn00868, cn01320, cn01319 and cn01747 are likely to be extracellular globins and *Hco-glb*_cn01356, cn08501, cn01583 and cn01377 are most likely to be intracellular.

After adjusting alignments 61 globin EST clusters were defined (Table 1A and Supplementary Table 1). From literature and database searches we also found thirteen sequences from previously published work (Table 1B). Combining the EST clusters, the four new sequences generated above and the sequences from the literature there are 72 distinguishable nematode globin sequences now identified when we exclude six EST clusters that are near identical to published sequences. Of the 72 putative genes, 62 had sufficient sequence to construct alignments and explore relationships between genes (Supplementary Table 2).

Evolution of nematode globin genes: Using nematode globin genes from published reports, our data for *T. canis* and *S. trachea*, and globins defined by EST clustering (Table 1 - excluding orphan sequences), we inferred a gene genealogy and compared this to published species trees constructed using small subunit ribosomal RNA (SSU rRNA) sequences (Meldal et al. 2007). As expected, due to a high degree of amino acid dissimilarity the consensus globin trees are not well resolved and many internal nodes are polytomous (Figure 1 and Supplementary figures 1 and 2). However, some clades are well supported in both the Bayesian analysis and by bootstrap analysis of the maximum parsimony (ML) and minimum evolution (ME) trees, and in some cases these groupings conflict with the family level structure of the published SSU rRNA species relationship tree (Meldal et al. 2007). In the Bayesian analysis, the most significant discrepancy between the SSU rRNA RNA tree and the globin tree is the separation of the majority of strongylid sequences into the clade marked "S" in Figure 1. This clade is separate from the major chromadorean clade and *Meloidogyne* spp. (Tylenchida) and *Caenorhabditis* spp. are shown as sister groups to the exclusion of all Strongylida. This conflicts with the SSU rRNA tree where Strongylida and Caenorhabdits are sister groups to the exclusion of Tylenchida; this may be a long-branch attraction artefact, and the group containing both Caenorhabditids and Tylenchids has low statistical support. In the ML and ME trees shown in Supplementary figures 1 and 2, the Caenorhabditid/Tylenchid clade has less than 50% bootstrap support and so is not shown.

Within the Strongylida, gene duplication has been rampant. There are multiple genes in all genera sampled through EST projects (*Ancylostoma*, *Haemonchus*, *Nippostrongylus*, *Ostertagia*, *Teladorsagia*). In some cases, there is an indication of homology between genes from separate taxa (*e.g.* clades M1 and M2 in Figure 1, which contain sequences from three trichostrongylid taxa with high bootstrap support). Some strongylid-derived protein sequences are significantly divergent from all other sequences in the group and form a separate clade (for example, GLB_ANCCA *cn01487 and* GLBM_SYNTR). These genes could define a separate family of globins perhaps lost in some strongylids such as *H. contortus*.

Correspondence of EST clusters with genes: The EST clusters we have generated may not correspond exactly to separate genes because of the complication introduced by sequencing errors and allelic variation. These sources of variation may contribute to a division of EST sequences, derived from the transcripts of a single gene, into more than one cluster. Our phylogenetic analysis provides evidence that some clusters are likely to be derived from separate genes when clusters from other taxa are consistently grouped as sister sequences to the exclusion of other sequences from the same taxon. The groups M1 and M2 described above are examples. In other instances, the correspondence of clusters to genes is more ambiguous (e.g. clade C1 in Figure 1 which contains three clusters from *H. contortus*). We undertook a number of investigations to try and establish whether the seven *H. contortus* predicted genes actually represent separate genes or are representative of extremely divergent alleles at fewer than seven loci.

First, alignment of the predicted position of SNPs (see below) within each cluster was compared to the alignment of the consensus nucleotide sequences of each cluster. If allelic variation had been the major determinant of division of sequences into multiple clusters, it would be expected that SNP positions would rarely correspond to divergent nucleotides between clusters. This was not the case; some divergent nucleotides corresponded to intra-cluster SNP predictions whilst others did not.

Second, alignment of cluster consensus sequences with raw sequence data from the incomplete *H. contortus* genome project called Sanger reads in figure 4A and below (http://www.sanger.ac.uk/Projects/H_contortus/) was undertaken to see if there was a correspondence of genomic loci and EST cluster sequences. For the four intracellular globin EST clusters, there was a significant match for only *Hco-glb*_cn08501 to genomic sequences (Sanger reads haem-1155c08.q1k, and haem-1033p03.q1k). Because the other three intracellular globin EST clusters display many differences to *Hco-glb*_cn08501 and are not matched to any Sanger read by Blastn, this implies that these four clusters must define at least two "real" genes. For the four extracellular globins, there was evidence of four divergent genes based on genomic sequence. No genomic sequence was a close match to *Hco-glb*_cn00868, whilst there were matches to *Hco-glb*_cn01320 (haem-70g17.q1k, haem-57j02.q1k, haem-1056p10.qlk), *Hco-glb*_cn01319 (haem-439c10.q1k, haem-804l05.q1k) and *Hco-glb*_cn01747 (haem-1075m01.q1k, haem-1099k13.p1k).

Divergent intron and 3'UTR sequences allowed alignment of *Hco-glb*_cn01320, *Hco-glb*_cn01319 and *Hco-glb*_cn01747 to separate genomic contigs, and those matching *Hco-glb*_cn01319 and *Hco-glb*_cn01320 were also used in phylogenetic analysis where they grouped separately (Figure 4A). This was an interesting observation as *Hco-glb*_cn01320 and *Hco-glb*_cn01319 share a high level of exonic nucleotide identity, but can be separated based on genomic sequence; implying that these "genes" have arisen from a recent duplication. We investigated this further by sequencing clones from both a cDNA library and from genomic DNA amplified using primers we designed to be specific for either *Hco-glb*_cn01320 or *Hco-glb*_cn01319. The sequence of the clones obtained was quite similar using either set of primers or template DNA, but amongst a high degree of apparent allelic variation (described below), the sequences obtained using the *Hco-glb*_cn01319 primers are distinct from those obtained using *Hco-glb*_cn01320 primers. One way of illustrating this is via phylogenetic analysis, and this is shown in Figure 4A.

Nematode globin gene structure: Genomic sequences are available for class I globins from *A. suum*, *B. malayi*, *C. elegans*, *Caenorhabditis briggsae*, *Caenorhabditis remanei*, *Mermis nigrescens*, *N. brasiliensis*, *P. pacificus*, *P. decipiens*, *S. trachea*, *T. canis* and *H. contortus*, and intron positions assigned (Table 2). The occurrence of conserved intron positions, relative to the predicted protein three-dimensional structure, in globin genes is well documented, with 5' and 3' introns at B12.2 and G7.0 occurring in many taxa across the plant and animal kingdoms, including in many nematode globins (Table 2). In the *Ppa-glb* genes (see Supplementary Figures 4-7) the predicted B helix is shorter than in the other globins, and in our amino acid alignment (Supplementary Table 2) residue B12 is represented by a gap. Because of this, the *Ppa-glb* introns 1 could be considered to be at B11.2. Central introns are found in nematodes and plants, and the placement of these is not conserved between these groups (Sherman et al. 1992; Moens et al. 1992). In nematodes, the central intron may be at E8.1 (ascaridid nematodes; Sherman et al. 1992) and we observed this intron position in both *T. canis* genes described here. The *B. malayi* globin (see Supplementary Figure 3) also probably has an intron at E8.1, but the intron donor sequence at this position is not optimal, and the alternative position of E9.0 cannot be ruled out. In other nematode clades the second intron may be located at E3.2, as in *Caenorhabditis* spp. and Strongylida (Blaxter et al 1994a; Kloek et al. 1996) and we observed this intron in the *S. trachea* and *H. contortus* globin genes described here and the *P. pacificus* genes predicted from genome sequence also have introns at E3.2. Intron loss is common in nematode globins, whereas it is relatively rarely observed in globins from chordates and angiosperms where divergent gene families from distantly related species often retain the same gene structure (Fuchs et al. 2005; Hunt et al.

TABLE 2. Intron positions in Class I globin genes from nematodes.

Gene Name	Intron position*								Reference
	B12.2	E3.2	E8.1	ef4b.0	ef4d.0	G7.0	H14.2	Inter-domain introns **	
As-glbm	no	no	no	no	no	no	no	no	Blaxter, et al., 2008
As-glbp GLBA	yes	no	yes	no	no	yes	no	yes	Sherman, et al., 1992
As-glbp GLBB	yes	no	yes	no	no	yes	no	yes	
Bma-glb ***	yes	no	yes	no	no	yes	no	no	Hoogewijs, et al., 2008
Cbr-glb	no	yes	no	no	no	no	no	no	Kloek et al., 1996
Ce-glb-1	no	yes	no	no	no	no	no	no	Kloek et al., 1996
Cre-glb	no	yes	no	no	no	no	no	no	Kloek et al., 1996
Mni-glbm	yes	no	no	no	no	yes	no	no	Burr, et al., 2000
Mni-glbe	yes	no	no	no	no	yes	no	no	Burr, et al., 2000
Nbr-glbm	yes	yes	no	no	no	yes	no	no	Blaxter et al., 1994a
Nbr-glbc	yes	yes	no	no	no	yes	no	yes	Blaxter et al., 1994a
Pde-glbp GLBA	yes	no	yes	no	no	yes	no	yes	Dixon et al., 1992
Pde-glbp GLBB	yes	no	no	no	no	no	no	yes	
Str-glbm	yes	yes	no	yes	no	yes	no	no	This report
Str-glbc	yes	yes	no	yes	no	yes	no	yes	This report
Tca-glbp	yes	no	yes	no	no	yes	no	yes	This report
Tca-glbm	yes	no	yes	no	no	yes	no	no	This report
*Hco-glb*_cn01320	yes	yes	no	yes	no	yes	no	yes	This report
*Hco-glb*_cn01319	yes	yes	no	yes	no	yes	no	yes	This report
*Hco-glb*_cn01747	yes	yes	no	yes	no	yes	no	yes	This report
*Hco-glb*_cn08501	yes	yes	no	yes	no	yes	no	no	This report
Ppa-glb ****	yes	yes	no	no	yes	yes	yes	no	This report

* - "yes" is reported where the intron is present in the gene sequence, "no" is reported where it is not. Two additional intron positions are known for *P. pacificus* globin which is not shown in the table (see text).

** - introns separating the secretory leader peptide-encoding domain from the globin domain, and/or introns separating adjacent globin domains (GLBP_ASCSU and GLBP_PSEDE).

*** - see text regarding *Bma-glb* intron positions; E9.0 and E8.1 cannot be distinguished given the available information.

**** - All four *Ppa-glb* genes have the same predicted gene structure.

2001). *C. elegans* has no 3' and 5' introns, retaining the central intron alone (Kloek et al. 1996), whereas in *A. suum* myoglobin there are no introns (Blaxter et al. 1994b). Both *S. trachea* globins and four predicted *H. contortus* globins (*Hco-glb*_cn01320, *Hco-glb*_cn01319, *Hco-glb*_cn01747, *Hco-glb*_cn08501) have an additional intron at position ef4b.0, giving them four introns in total (see methods for description of the modified nomenclature). The four predicted *P. pacificus* globins have five introns, with additional introns at positions ef4d.0 and H14.2. The ef4d.0 intron in *Ppa-glb* genes does not appear to be homologous to the ef4b.0 intron from Strongylids. These differing positions are not likely to be artefacts of ambiguous amino acid alignment as the ef interhelix region is quite conserved between these groups (Figure 2). Combined with the observation of introns between secretory leader encoding sequences and the main globin domain in a variety of species, and inter-domain introns in the di-domain globins from *A. suum* and *P. decipiens*, there have clearly been multiple examples of intron insertion in nematode globin genes.

Our assigned intron positions for *Bma-glbm* and *Ppa-glb* genes differ from those published previously (Hoogewijs, 2008) because we have revised the gene structure predictions of these genes in the *B. malayi* and *P. pacificus* genomes and because of differences in our alignments of *P. catadon* myoglobin with the nematode globins. Irrespective of the exact intron positions, the *P. pacificus* genes have additional introns in unique positions for class I nematode globins, further indicating that the late acquisition of introns has occurred in this gene family.

Allelic variation in nematode globins (Predicted SNPs from EST clusters): Twenty five EST clusters were analysed, each of which comprised 5 or more EST sequences representing separate clones from cDNA libraries. These clusters were from 16 species of nematodes. Table 1 shows the number of predicted SNPs from these clusters according to the methodology described herein. The number of predicted SNPs varies greatly among the clusters, from 0 to 66. Clusters from trichostrongylid nematodes had a greater number of predicted SNPs than other groups, with an average of 1.72 SNPs per EST, compared to 0.95 over the whole dataset and 0.14 for the Spiruria. Although this undoubtedly reflects a greater level of genetic diversity within species in the Trichostrongylida, it is also conditioned by the number of individuals sampled to make the cDNA libraries that were used to generate the EST data. *A. suum*, *A. lumbricoides*, *T. canis* and *Onchocerca volvulus* are all large nematodes as adults (>5 cm), and libraries created from these species were probably constructed using fewer individuals. Therefore, the capacity to capture genetic diversity in these libraries was limited, leading to the observed lower amount of variation in these globin gene clusters. *C. elegans* cDNA libraries are made from many thousands of individuals, but these are from highly inbred, genetically homozygous lines, and we did not find any predicted SNPs by analysing EST sequences which aligned to the *Ce-glb-1*.

Figure 3 shows the number of observed biallelic SNPs divided into transitions (Ts) and transversions (Tv). As would be expected, the number of Ts exceeds the number of Tv across the data set (mean Ts/Tv = 3.09). For individual clusters however, Ts/Tv varies from 1 to 9.3; for the trichostrongylid group, this ratio is close to the mean across the data set (3.21), however for the spirurine group, there is a relative excess of Tv resulting in a combined Ts/Tv of 1.17. This is an interesting finding, as the incidence of Tv mutations in nature is usually documented as being far lower relative to Ts.

Allelic variation in nematode globins (polymorphism in T. canis *globins):* Twelve clones of *Tca-glbm* were generated by RT-PCR from two separate RNA pools, from a total of five animals (or 10 haploid chromosome sets). Gonadal tissue was excluded by dissection. Consistent variability was observed between sequences, and we propose that five alleles were identified of a possible 16 which differ bi-allelically at four nucleotide positions. The variant nucleotides are a silent A to G transition at 3rd position ab2 (*A. suum* D1 numbering), a silent C to T transition at 3rd position H2, and two adjacent non-synonymous, C to T transitions at the 1st and 2nd positions of a codon in the C-terminal region of the predicted protein which cannot be aligned with the *A. suum* D1 structure. These non-synonymous SNPs could cause this codon to encode serine (TCT), leucine (CTT), proline (CCT) or phenylalanine (TTT), depending on the sequence, however only TCT and CTT sequences have been observed in the clones we sequenced.

The gene structures of both *Tca-glbm* and *Tca-glbp* have been elucidated. Sequencing of five clones of a region containing intron 4 from *Tca-glbp* identified two alleles with different length introns (Figure 4B). The polymorphism is complex and may have arisen following either two deletion events in one allele (represented by 2

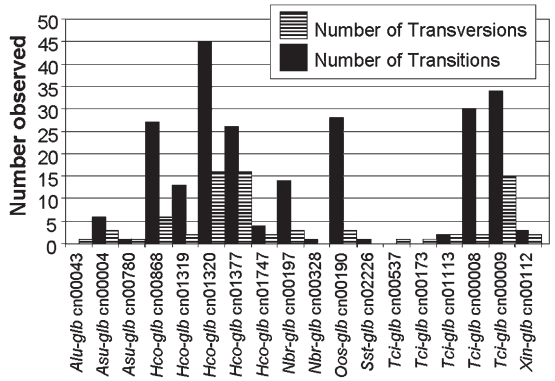
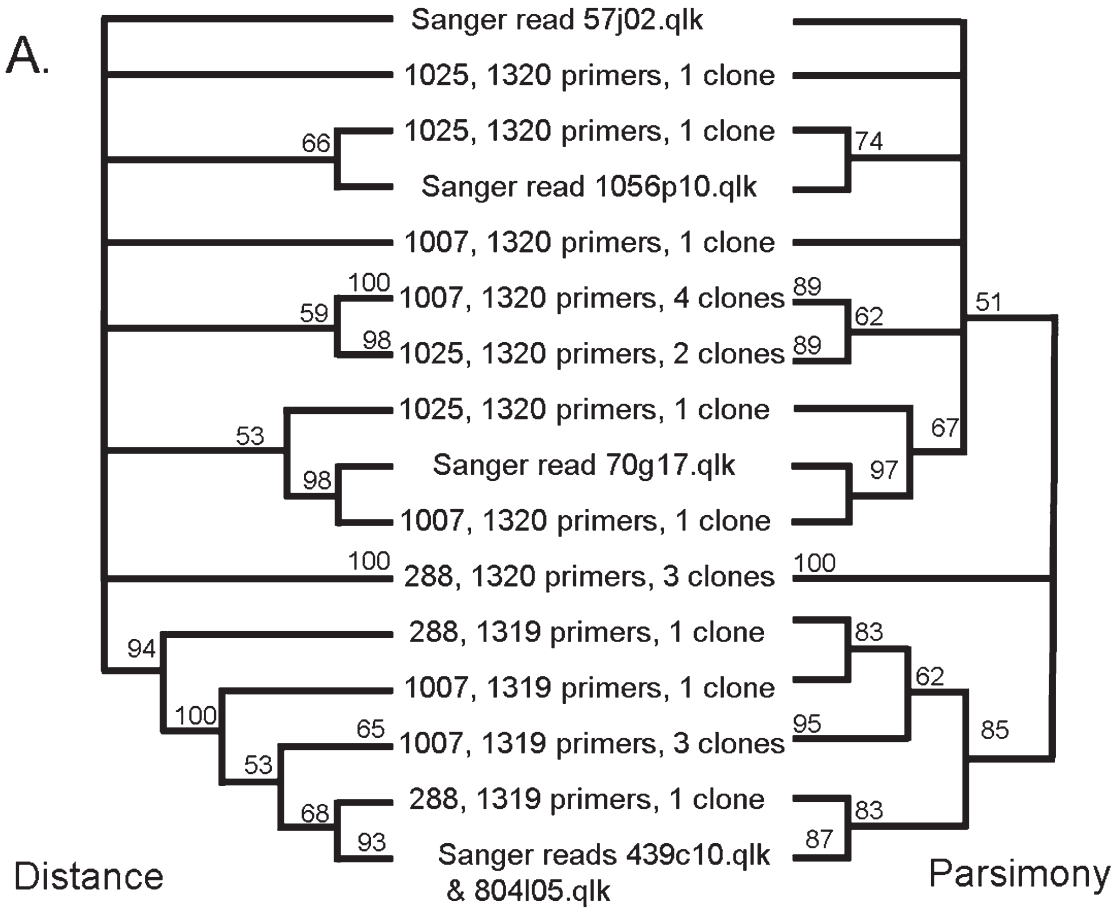


FIG. 3. The number of predicted SNPs from 24 globin EST clusters, illustrating how the relative abundance of genetic variation differs between genes and between species. Transition and transversion SNPs are shown in separate columns for comparison.



B.

```
intron 4Δ CAA AAG GAT GAA TTC TTT GCG AAA CAA G gt.............43 bp deletion.................ttccgcctttgctact
intron 4  CAA AAG GAT GAA TTC TTT GCG AAA CAA G gtattcttttccgccttcggtactccatgaatkctcgattcaaaatt.cgccttcggtact 801
           Q   K   D   E   F   F   A   K   Q

intron 4Δ ccatggatkctcgattcaatataactgagttaaaagaancgtggcaagaacgaaggaagaaaccgttgaacttctttctccctctctcctctccaaa
intron 4  ccatgaattctcgattcaatataactgagttgaaagaagcgtggcaagaacgaangaagaaaccgttgaacttctttctccctctctcctctccaaa 901

intron 4Δ cctcgtcatccctcaattgtcgaaggcgcaaaatttgcaacacca.............52 bp deletion.....................tt
intron 4  ccccgccgtccttcaattatcgaaggcgcaaaatttgcaacaccgttattaatttgcaacatcaaaatttgcaacaacgttattatggtcgcaaaattt 1001

intron 4Δ gcgatcataataacgttnacaacatcttagactgttttgttttgctagaataaggcgcatacnatctatggaggacgggaaaactggaacatttaaatg
intron 4  gcgatcataataacgttaaaaacatcttagactgttttgttttactagaataaagcgcatacaatctatggaggaggggaaaacctgaagatttaaatg 1101

intron 4Δ aaattcgccnataagcattttctctttccagaagaaacgtttcttggaatttccatgccctatcatgataacaacaattatgatatgataacaataatg
intron 4  agttttgccaataagcattttctctttccagaagaaacgtttcttggaatttccatgccctattatgataacaacaattatgatatgataacaacaatg 1201

intron 4Δ ataacaatgacngcaatagccaaagcataaggtattgcgcatttttatgcgattttcctaagactgccaattcatcaaacacccctatcgtagtcgaacg
intron 4  ataacaatgacngcaatagccaaagcataaggtattgcgcatttttatgcgattttcctaagactgccaattcatcaaacacccctatcgtagtcgaacg 1301

intron 4Δ aaagttcttcggcgaacttgagcttgatccatttttcttgtgagtttcgcttcttcttgcacttctttgaataaataatctaatcgaatcagcacatatt
intron 4  aaagttcttcggcgaacttgagcttgatccatttttcttgtgagtttcgcttcttcttgcacttctttgaataaataatctaatcgaatcagcacatatt 1401

intron 4Δ aaccatttttctaaacttcaagaactcgagggcattttcttaagtaagccataaaggagattgtaatcccaagaatgtgaacaaaaaaaaacaaaaatc
intron 4  aaccatttttctaaacttcaagaactcgagggcattttcttaagtaagccataaaggagattgtaatcccaagaatgtgaacaaaaaaaaacaaaaatc 1501

intron 4Δ aaaattgcatcgaacaattgttttcagtttttccacacattcgataattctatgagcaagagaagatgtttctcttagacagaattcgcactcatcaca
intron 4  aaaattgcatcgaacaattgttttcagtttttccacacattcgataattctatgagcaagagaagatgtttctcttagacagaattcgcactcatcaca 1601

intron 4Δ ttcatatgctcgaattttatttccatgtacttaacgacgatataattttgcgcctaatatacag GT CAA AAA ATC CTG CTC GCC TGT CAC
intron 4  ttcatatgctcgaattttatttccatgtacttaacgacgatataattttgcgcctaatatacag GT CAA AAA ATC CTG CTC GCC TGT CAC 1691
                                                                        G   Q   K   I   L   L   A   C   H
```

FIG. 4. Allelic variation in nematode globins. (A) Clones derived from a genomic DNA fragment of *Hco-glb_*cn*01319* or *Hco-glb_*cn*01320* were sequenced from either 3 individuals (288, 1025 and 1007 – cn01320) or 2 individuals (1025 and 1007 – cn01319). These were aligned with sequence reads fromt the *H. contortus* genome sequencing project, and a phylogenetic analysis undertaken. The relationships between clones are illustrated by cladograms with the result of minimum evolutionary distance to the left "Distance" and maximum parsimony to the right "Parsimony". The results suggest that cn01320 and cn01319 are separate genes as these group separately in the tree. (B) Comparison of two alleles of *Tca-glbp* that differ in the length of intron 4. A number of SNPs can also be seen.

clones) or by two insertion events in the other (3 clones). In comparison to the EST sequence data, one predicted SNP (TC00537) was in a position distinct from the five SNPs observed by multiple sequencing of *Tca-glbm* cDNAs (data not shown).

Allelic variation in nematode globins (polymorphism in a H. contortus *globin):* Our *in silico* prediction of SNPs based on EST alignment revealed 43 synonymous and 27 non-synonymous SNPs for *Hco-glb*_cn01320. Three indels causing premature stop codons presumably leading to truncated translation were also found. The non-synonymous SNPs caused either amino acid changes (24) or premature stop codons (3). When aligned with GLBP_ASCSU, for which the 3 dimensional structure has been partially resolved (Yang et al. 1995), the amino acid changes are predicted in a number of helices and inter-helix regions as described in Table 3. When we sequenced a portion of the gene (exons 3, 4 and 5 and a small portion of exon 2) from 3 individuals, we again observed 21 of the 33 synonymous SNPs and 5 of the 7 non-synonymous SNPs that had been predicted from EST analysis. An additional 11 synonymous and 1 non-synonymous SNPs were observed that had not been predicted from EST sequence alignment, and a further 66 SNPs and 2 insertion-deletion mutations were observed over three intron regions. The residues affected by the non-synonymous SNPs were in the ef loop, helix F10, H11 and two in parts of the alignment C-terminal of the H-helix aligned with the GLBP_ASCSU crystal structure (Table 3).

Gene expression patterns of T. canis *globins:* Both *Tca-glbm* and *Tca-glbp* were initially cloned by RT-PCR using RNA derived from whole adult female nematodes. Further analysis of expression using tissues dissected from live adult females and from male RNA indicated that both globins were expressed in both sexes (Figure 5A, B). The tissue distribution of expression differed between isoforms; *Tca-glbp* is expressed in the female reproductive tract and both *Tca-glbp* and *Tca-glbm* are expressed in the body wall of adult females, in males and in the head region of females. No expression of either gene was evident from intestinal tissue.

Gene expression patterns for nematode globins derived from EST data: The cDNA libraries from which the publicly

TABLE 3. Details of non-synonymous SNPs and coding region indels predicted in *Hco-glb*_cn01320.

Gene Region[1]	SNP or indel	AA change	3D position[2]	Present in Re-sequencing[3]	minor allele frequency[4]	number of ESTs
Exon0	A/G	Truncated translation	SLP	NA	0.33	6
Exon0	G/T	Truncated translation	SLP	NA	0.33	6
Exon0	G/T	Q-H	SLP	NA	0.17	12
Exon0	A/T	L-H	SLP	NA	0.12	17
Exon0	A/C	I-L	SLP	NA	0.39	18
Exon0	C/T	L-F	SLP	NA	0.20	20
Exon0	G/T	V-F	SLP	NA	0.13	24
Exon0	G/T	F-C	SLP	NA	0.08	25
Exon0	G/T	V-G	SLP	NA	0.08	25
Exon0	-/C	Truncated translation	SLP	NA	0.18	22
Exon0	A/T	F-Y	SLP	NA	0.16	25
Exon0	G/C	F-L	SLP	NA	0.14	21
Exon0	C/T	T-I	SLP	NA	0.16	25
Exon1	-/T	Truncated translation	A1	NA	0.15	26
Exon1	G/T	D-Y	A2	NA	0.07	27
Exon1	-/A	Truncated translation	A12	NA	0.11	28
Exon2	C/T	R-C	C5	NA	0.11	27
Exon2	G/T	R-L	C5	NA	0.17	29
Exon2	A/T	Truncated translation	C6	NA	0.14	29
Exon2	A/T	K-I	C6	NA	0.17	29
Exon2	A/T	K-N	C6	NA	0.11	28
Exon2	C/G	S-R	cd loop	NA	0.07	29
Exon2	A/G	A-T	D2	NA	0.07	29
Exon3	A/T	F-Y	ef loop	Yes	0.35	26
Exon4	G/T	S-A	F10	Yes	0.36	25
Exon5	C/T	P-S	gh loop	No	0.08	24
Exon5	A/G	A-T	H6	No	0.13	24
Exon5	A/G	I-V	H11	Yes	0.13	24
Exon5	A/G	T-A	post H	Yes	0.26	23
Exon5	C/G	T-S	post H	Yes	0.09	23

1 - "Exon0" represents the part of this predicted gene encoding the secretory leader peptide (SLP). In other extracellular globins there is an intron separating this from the remainder of the gene, however we have no evidence from genome sequencing for this intron in *Hco-glb*_cn1320.

2 - SLP residues are those predicted to be part of the secretory leader peptide, "post H" refers to residues C-terminal of the H-helix aligned with the *GLBP ASCSU* crystal structure, and the other designations refer to positions aligned with the *GLBP ASCSU* crystal structure.

3 - NA for this table is for "not attempted". Re-sequencing was conducted using a partial gene fragment.

4 - The minor allele frequency was calculated by dividing the number of EST sequences containing the rarer allele by the number of EST sequences in the alignment at the position indicated. Sequences containing a residue or gap not represented by the two most common alleles were excluded.

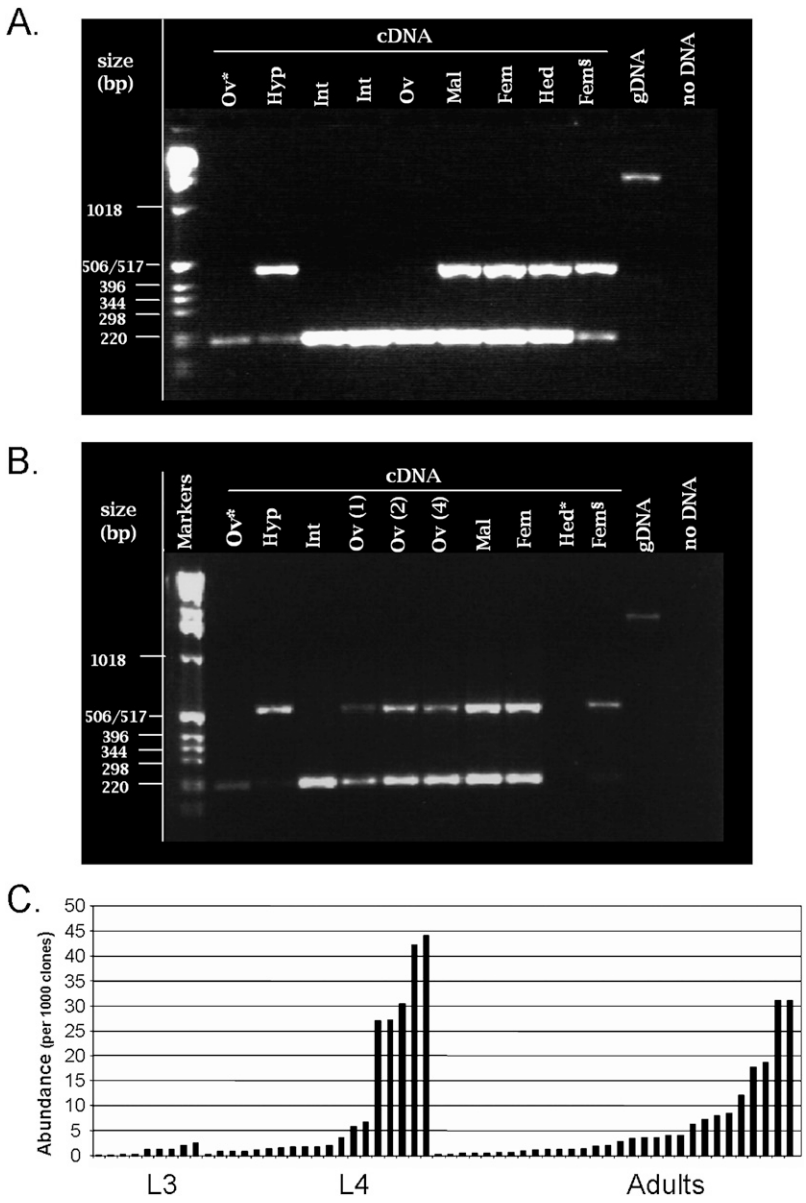


FIG. 5. Gene expression of nematode globins. *Tca-glbm* (A) and *Tca-glbp* (B) in various tissues of *T. canis*. Reverse transcriptase PCR was undertaken in a multiplex reaction, co-amplifying mRNA for *Tca-rpl-3* as an internal control. Lane one shows a DNA size standard with fragment sizes indicated, and the remaining lanes contain PCR reactions. Template is cDNA reverse transcribed with d(T) primer from Ovary (Ov), Hypodermis and longitudinal muscle (Hyp), Intestine (Int), male nematode (Mal), Female whole nematode (Fem), the head section of a female nematode, anterior to the junction of the pharyngeal basal bulb and intestine (Hed) or *T. canis* genomic DNA (gDNA) or no template control (noDNA). Both *Tca-glbm* and *Tca-glbp* amplicons are approximately 500 base pairs in length, whereas the *Tca-rpl-3* amplicon is 220 base pairs. Ov(1), Ov(2) and Ov(4) show a 2-fold dilution series of template. (C) The abundance of Strongylid globins in EST sequences is an indication of the level of their expression in the tissues from which cDNA libraries were obtained. The graph shows the relative abundance of 58 cluster-predicted globins from 7 Strongylid nematode species (*T. circumcincta*, *O. ostertagi*, *H. contortus*, *N. americanus*, *A. caninum* and *A. ceylanicum*), and it is clear that expression is higher in parasitic L_4 and adult stages compared to the infective L_3 stage.

available globin ESTs were derived are from various life cycle stages, tissues and other treatments. Supplementary Table 3 shows expression of the nematode globins with details of the cDNA libraries from which they were derived. Some estimates of globin expression in these nematodes and nematode tissues can be obtained by dividing the number of globin clones by the number of sequenced clones from the library. In this way, we predict that strongylid nematodes and *Ascaris* spp. express globins to the highest levels in the dataset (Table 4). Plant parasitic nematodes and free-living nematodes have lower levels of expression.

The *A. suum* dataset includes libraries constructed using cDNA from dissected tissues; and so a comparison with our qPCR analysis of *T. canis* globins can be made. As with our *T. canis* results, the intracellular globin is most highly expressed in the hypodermis/body wall muscle and head tissues. The *Ascaris* library results also show that the L_4 larval stage expresses *As-glb*_cn00004 to a high level. Considering the high abundance of

TABLE 4. The top 15 nematode cDNA libraries with the highest relative level of globin expression by cluster.

Rank	Library name	Species	Cluster	Abundance (per 1000 clones)
1	L_4 SL1 from David Knox *Haemonchus contortus*	*H. contortus*	cn00868	44.04
2	L_4 SL1 from David Knox *Haemonchus contortus*	*H. contortus*	cn01377	42.20
3	adults *Teladorsagia circumcincta*	*T. circumcincta*	cn00008	31.06
4	adults *Teladorsagia circumcincta*	*T. circumcincta*	cn01113	31.06
5	L_4 library *Teladorsagia circumcincta*	*T. circumcincta*	cn00008	30.46
6	L_4 library *Teladorsagia circumcincta*	*T. circumcincta*	cn00009	27.07
7	L_4 SL1 TOPO v1 *Ostertagia ostertagi*	*O. ostertagi*	cn00190	26.96
8	L_4 pSPORT1 Zarlenga v1 *Ascaris suum*	*A. suum*	cn00004	22.86
9	adult male whole body *Ascaris lumbricoides*	*A. lumbricoides*	cn00043	20.31
10	adults *Teladorsagia circumcincta*	*T. circumcincta*	cn00173	18.63
11	uni-zap adult library *Nippostrongylus brasiliensis*	*N. brasiliensis*	cn00197	17.66
12	(parasitic nematode) body wall muscle and hypodermis *Ascaris*	*A. suum*	cn00004	15.87
13	adult (full length enriched) *Nippostrongylus brasiliensis*	*N. brasiliensis*	cn00197	12.00
14	female head SL1 TOPO v1 Murphy Chiapelli McCarter	*A. suum*	cn00004	11.82
15	male head SL1 TOPO v1 Murphy Chiapelli McCarter	*A. suum*	cn00004	10.09

protein, it is surprising that the di-domain extracellular globin from *A. suum* (*As-glb*_cn00019) has an apparent mRNA abundance lower than for the intracellular globin. In contrast to *Tca-glbp*, *As-glb*_cn00019 is not highly expressed in the female gonad, but rather has a very similar pattern of expression to *As-glb*_cn00004; suggesting that the functions of *Tca-glbp* and the di-domain extracellular globin from *A. suum* (*As-glbp*) are divergent. Another *A. suum* cluster, *As-glb*_cn00780, appears to encode a single domain globin from *A. suum* that has higher sequence similarity to *Tca-glbp* than *As-glb*_cn00019; it is interesting that cDNAs from this cluster are predominantly from libraries obtained from gut and ovarian tissue, which is a pattern dissimilar to either *Tca-glbm* or *Tca-glbp*. *A. suum* also had a predicted second intracellular globin (*As-glb*_cn17423), which is expressed at a very low level, the sequenced clones are both from a female gut-derived library.

The majority of cDNA libraries with a high abundance of globin cDNA clones were from strongylid nematodes. There were 8 clusters of ESTs from four strongylid nematode species (*H. contortus*, *T. circumcincta*, *O. ostertagi* and *N. brasiliensis*) that had globin clone abundances above 1%. All these libraries were derived from L_4 or adult stage nematodes. When compared to L_3 or egg derived libraries, it appears that globin expression in strongylids is enhanced during parasitic life stages (Figure 5C). Considering *H. contortus* alone, no globin ESTs were detected in L_3 or egg libraries, whilst abundances in an SL1 trans-spliced L_4 library ranged from 3.7 per 1000 for *Hco-glb*_cn01356 to 44.0 for *Hco-glb*_cn00868. In two adult libraries abundances ranged from 0.43 per 1000 in one library to 8.4 in a day 11 adult library for *Hco-glb*_cn01320. Interestingly, *Hco-glb* n00868 (extracellular), *Hco-glb*_cn01356 (intracellular) and *Hco-glb*_cn01377 (intracellular) are all highly expressed in L_4 and appear to be L_4 specific; no other *H. contortus* clusters contained EST sequences from the L_4 stage library.

DISCUSSION

Globins are almost ubiquitous genes found in the genomes of organisms from all four eukaryotic kingdoms, Eubacteria and Archaea (Vinogradov et al. 2006). Nematodes have been known to express globins for some time (Adduco 1889; Blaxter 1993; Frenkel et al. 1992), and they have been demonstrated to have divergent gene structures and functions (Blaxter et al. 1994b; Burr et al. 2000; Kloek et al. 1996; Moens et al. 1992; Sherman et al. 1992; Vinogradov and Moens, 2008). In the work described here we have shown that nematode globin genes display a great deal of sequence and gene structure diversity. There is evidence for both recent and ancient globin gene duplications, especially in the Strongylid and Ascarid groups. The presence of introns in a fourth, "non-standard" position in Strongylid globin genes from *S. trachea* and *H. contortus*, and of other additional introns in the *P. pacificus* globin genes show that introns have been inserted subsequent to the putative assembly of a protoglobin gene from gene fragments in the earliest eukaryotes (Gilbert et al. 1986). These introns are not associated with additions of new structural domains in contrast to the introns separating secretory leader peptide domains or globin domains of di-domain globins. Clearly some globin introns can be "late" in addition to others being "early". In *C. elegans* a larger family of globin like genes have been identified (Vinogradov et al. 2006; Hoogewijs et al. 2007). The globins discussed above are most similar to the *C. elegans* globin ZK637.13 defined as Class I globins in Hoogewijs, et al., 2008. The gene structures exhibited by the other Class II globins and globin-like genes are also divergent from the conserved pattern observed in globins from Chordates and other groups (Hoogewijs, et al., 2008).

The function of globins in invertebrate animals remains enigmatic (Vinogradov and Moens, 2008).

Although there have been some illuminating investigations (Burr et al. 2000; Kimura et al. 1999; Minning et al. 1999), these suggest many divergent functions rather than a single, global purpose for these proteins.

Within the nematode taxa considered here, we have shown that the pattern of expression of globins follows some consistent patterns. Expression in parasitic species seems to be increased in adult and pre-adult stages, though these do not necessarily correspond to parasitic stages, as in *M. nigrescens* (Burr et al. 2000). This pattern is not conserved in *C. elegans* however where *Ce-glb-1* is expressed at roughly equivalent levels in the L_3, adult and alternate L_3 dauer stage, and is also expressed in eggs, L_1 and L_2 (Hoogewijs, et al., 2007). It is also of interest that *Ce-glb-1* gene expression does not respond to hypoxia as other globins do in plants and other invertebrate animals (e.g. Kimura, et al., 1999; Hunt, et al., 2001), but is instead responsive to perturbations in insulin signalling (Hoogewijs, et al., 2007). The parasitic nematode gene expression patterns may indicate divergence from *Ce-glb-1* if the main stimulus for increased expression is the movement from normoxic to hypoxic environments, however there are a whole range of factors encountered by parasitic larvae on their entrance to the host, so changes in insulin signalling could well be occurring and influencing gene expression at this stage. This is an area of investigation worth pursuing in future work.

Expression of globins in the hypodermis also seems to be a common observation for Class I nematode globins. In contrast, the Class II globins described are expressed predominantly in neurons (Hoogewijs, et al., 2008), including one from *C. elegans* (*glb*-5) which has been shown to be important for the avoidance of high oxygen concentrations (Persson, et al., 2009; McGrath, et al., 2009).

Other aspects of the investigation we have undertaken support the hypothesis of a diversity of functions in nematode globins. First we note that many nematode globins have putative secretory leader peptides and therefore have roles outside the cell rather than internally whilst other globins are clearly located intracellularly. Second, there is significant sequence divergence in the genes we have identified and many globins have divergent residues at the E7 ligand coordinating position. Glutamine and leucine are the most common E7 residues observed in our alignment (Figure 2), and these residues have both divergent sidechain properties and have been shown to impart divergent ligand binding kinetics to globins when substituted for one another (Hargrove et al. 1996). Third, we have observed a high degree of genetic diversity between globins from different species and between multiple globin genes within species. This variation is highly suggestive of actively evolving proteins being selected for divergent function. It seems very likely that the further investigation of these proteins will reveal a diversity of both biochemical and cellular functions which may be useful for biotechnological applications or as targets for the control of parasites through immunological or pharmaceutical means.

The impressive degree of gene diversity we have observed is almost matched by the level of within-gene allelic diversity. The abundance of SNP in the *Hco-glb*_cn01320 gene (137 SNP per 1000 bp in exons and UTRs, and far higher in intron sequence) far exceeds that observed in a large segment of the *C. elegans* genome (1.1 SNP per 1000 bp (Wicks et al. 2001)). The abundance of putative SNP in most globin genes we analysed using *in silico* methods was also high, except for the *C. elegans* globin gene (ZK637.13) because the ESTs we used were all obtained from cDNA originating from the laboratory strain N2. Estimates of nucleotide diversity in coding regions of nematode genes have been attempted for a variety of nematodes, and these range from 5.8/1000 for wild *C. elegans* (Cutter 2006) to 68/1000 for both the obligate out-crossing *C. remanei* (Cutter et al. 2006) and 20/1000 for the mean *in silico* predicted SNP from 1,548 EST clusters in our *H. contortus* database WormSIS (data not shown). Therefore the available evidence suggests allelic diversity in some globin-encoding genes is higher than that exhibited by other nematode genes.

This observed allelic diversity has two important consequences. Firstly for evolutionary studies, these genes are not appropriate for studies of nematode taxon relationships. The rate of evolution is far too high for meaningful comparisons. We have also tried to use *Hco-glb*_cn01320 for population-level studies with little success, mostly because of the high level of diversity within isolates (data not presented). Secondly, the use of globins as vaccine or drug targets could be problematic. Although use of these genes as a vaccine target has met with limited success allelic divergence could have affected the consistency of results (Frenkel et al. 1992; Claerebout, et al., 2005; S. McClure and D. Emery, personal communication). A second consideration for this application would be the observation of multiple genes in economically important species, perhaps necessitating the use of multivalent approaches.

In conclusion, we have demonstrated that nematode globins show a huge amount of diversity in sequence and gene structure, though the timing and tissue specificity of expression may be more highly conserved. There is evidence for multiple gene duplications, multiple intron insertions and losses and for allelic variation at both synonymous and non-synonymous sites. The gene family shows great promise for discovering both unique insights into both globin structure-function relationships and for cellular roles in an important animal phylum.

Literature Cited

Adducco, V. 1889. "La substance colorante rouge de l'Eustrongylus gigante." Archives Italiennes de Biologie 11:52–69.

Blaxter, M. L. 1993. Nemoglobins - divergent nematode globins. Parasitology Today 9:353–360.

Blaxter, M. L., Ingram, L., and Tweedie, S. 1994a. Sequence, expression and evolution of the globins of the parasitic nematode *Nippostrongylus brasiliensis*. Molecular and Biochemical Parasitology 68:1–14.

Blaxter, M. L., Vanfleteren, J. R., Xia, J., and Moens, L. 1994b. Structural characterization of an *Ascaris* myoglobin. Journal of Biological Chemistry 269:30181–30186.

Blaxter, 1996. Protein motifs in filarial chitinases. Parastology Today 12:42.

Blaxter, M. L., Guiliano, D. B., Scott, A. L., and Williams, S. A. 1997. A unified nomenclature for filarial genes. Parasitology Today 13: 416-417.

Burr, A. H. J., Hunt, P. W., Wagar, D. R., Dewilde, S., Blaxter, M. L., Vanfleteren, J. R., and Moens, L. 2000. A hemoglobin with an optical function. Journal of Biological Chemistry 275:4810–4815.

Claerebout, E., Smith, W. D., Pettit, D., Geldhof, P., Raes, S., Geurden, T., and Vercruysse, J. 2005. Protection studies with a globin-enriched protein fraction of *Ostertagia ostertagi*. Veterinary Parasitology 128:299–307.

Cutter, A. D. 2006. Nucleotide polymorphism and linkage disequilibrium in wild populations of the partial selfer *Caenorhabditis elegans*. Genetics 172:171–184.

Cutter, A. D., Baird, S. E., and Charlesworth, D. 2006. High nucleotide polymorphism and rapid decay of linkage disequilibrium in wild populations of *Caenorhabditis remanei*. Genetics 174:901–913.

De Baere, I., Liu, L., Moens, L., Van Beeumen, J., Gielens, C., Richelle, J.. Trottmann, C., Finch, J., Gerstein, M., and Perutz, M. 1992. Polar zipper sequence in the high-affinity hemoglobin of Ascaris suum: Amino acid sequence and structural interpretation. Proceedings of the National Academy of Sciences, USA 89: 4638–4642.

Dixon, B., Walker, B., Kimmins, W., and Pohajdak, B. 1991. Isolation and sequencing of a cDNA for an unusual hemoglobin from the parasitic nematode *Pseudoterranova decipiens*. Proceedings of the National Academy of Sciences U.S.A. 88:5655–9.

Dixon, B., Walker, B., Kimmins, W., and Pohajdak, B. 1992. A nematode hemoglobin gene contains an intron previously thought to be unique to plants. Journal of Molecular Evolution 35:131–136.

Emanuelsson, O., Brunak, S., von Heijne, G., and Nielsen, H. 2007. Locating proteins in the cell using TargetP, SignalP and related tools. Nature Protocols 2:953–971.

Frenkel, M. J., Dopheide, T. A. A., Wagland, B. M., and Ward, C. W. 1992. The isolation, characterization and cloning of a globin-like, host-protective antigen from the excretory-secretory products of *Trichostrongylus colubriformis*. Molecular and Biochemical Parasitology 50:27–36.

Fuchs, C., Luckhardt, A., Gerlach, F., Burmester, T., and Hankeln, T. 2005. Duplicated cytoglobin genes in teleost fishes. Biochemistry and Biophysics Research Communications 337:216–223.

Fuhrman, J. A., Lee, J., and Dalamagas, D. 1995. Structure and function of a family of chitinase isozymes from Brugian microfilariae. Experimental Parasitology 80:672–680.

Gilbert, W., Marchionni, M., and McKnight, G. 1986. On the antiquity of introns. Cell 46:151–153.

Hargrove, M. S., Barrick, D., and Olson, J. 1996. The association rate constant for heme binding to globin is independent of protein structure. Biochemistry 35:11293–11299.

Hoogewijs, D., Geuens, E., Dewilde, S., Vierstraete, A., Moens, L., Vinogradov, S. N., and Vanfleteren, J. R. 2007. Wide diversity in structure and expression profiles among members of the Caenorhabditis elegans globin protein family. BMC Genomics 8:356–374.

Hoogewijs, D., De Henau, S., Dewilde, S., Moens, L., Couvreur, M., Borgonie, G., Vinogradov, S. N., Roy, S. W., and Vanfleteren, J. R. 2008. The Caenorhabditis globin family reveals extensive nematode-specific radiation and diversification. BMC Evolutionary Biology 8:279–300.

Hunt, P. W., Knox, M. R., Le Jambre, L. F., McNally, J., and Anderson, L. J. 2008. Genetic and phenotypic differences between isolates of *Haemonchus contortus* in Australia. International Journal for Parasitology 38:885–900.

Hunt, P. W., Watts, R. A., Trevaskis, B., Llewellyn, D. J., Burnell, J., Dennis, E. S., and Peacock, W. J. 2001. Expression and evolution of functionally distinct hemoglobin genes in plants. Plant Molecular Biology 47:677–692.

Kimura, S., Tokishita, S., Ohta, T., Kobayashi, M., and Yamagata, H. 1999. Heterogeneity and differential expression under hypoxia of two-domain hemoglobin chains in the water flea, *Daphnia magna*. Journal of Biological Chemistry 274:10649–10653.

Kloek, A., McCarter, J., Setterquist, R., Schedl, T., and Goldberg, D. 1996. *Caenorhabditis* Globin genes: Rapid intronic divergence contrasts with conservation of silent exonic sites. Journal of Molecular Evolution 43:101–108.

McGrath, P. T., Rockman, M. V., Zimmer, M., Jang, H., Macosko, E. Z., Kruglyak, L., and Bargmann, C. I. 2009. Quantitative mapping of a digenic behavioral trait implicates globin variation in C. elegans sensory behaviors. Neuron 61:692–699.

Meldal, B. H. M., Debenham, N. J., De Ley, P., De Ley, I. T., Vanfleteren, J. R., Vierstraete, A. R., Bert, W., Borgonie, G., Moens, T., Tyler, P. A., Austen, M. C., Blaxter, M. L., Rogers, A. D., and Lambshead, P. J. 2007. An improved molecular phylogeny of the Nematoda with special emphasis on marine taxa. Molecular Phylogenetics and Evolution 42:622–636.

Minning, D. M., Gow, A. J., Bonaventura, J., Braun, R., Dewhirst, M., Goldberg, D. E., and Stamler, J. 1999. *Ascaris* haemoglobin is a nitric oxide-activated 'deoxygenase'. Nature 401:497–502.

Moens, L., Vanfleteren, J., De Baere, I., Jellie, A. M., Tate, W., and Trotman, C. N. 1992. Unexpected intron location in non-vertebrate globin genes. F.E.B.S. Letters 312:105–109.

Parkinson, J., Whitton, C., Schmid, R., Thomson, M., and Blaxter, M. L. 2004. NEMBASE: a resource for parasitic nematode ESTs. Nucleic Acids Research 32:D427–430.

Persson, A., Gross, E., Laurent, P., Busch, K. E., Bretes, H., and de Bono, M. 2009. Natural variation in a neural globin tunes oxygen sensing in wild *Caenorhabditis elegans*. Nature 458:1030–1035.

Ronquist, F., and Huelsenbeck, J. P. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 19:1572–1574.

Sherman, D. R., Kloek, A. P., Krishnan, B. R., Guinn, B., and Goldberg, D. E. 1992. *Ascaris* hemoglobin gene: plant-like structure reflects the ancestral globin gene. Proceedings of the National Academy of Sciences U.S.A. 89:11696–11700.

Swofford, D. L. 2000. PAUP* Phylogenetic analysis using parsimony * and other methods. New York: Sinauer Associates.

Tweedie, S., Grigg, M. E., Ingram, L., and Selkirk, M. E. 1993. The expression of a small heat shock protein homologue is developmentally regulated in *Nippostrongylus brasiliensis*. Molecular and Biochemical Parasitology 61:149–154.

Vanfleteren, J. R., Van de Peer, Y., Blaxter, M. L., Tweedie, S. A., Trotman, C., Lu, L., Van Hauwaert, M. L., and Moens, L. 1994. Molecular genealogy of some nematode taxa as based on cytochrome c and globin amino acid sequences. Molecular Phylogenetics and Evolution 3:92–101.

Vercauteren, I., Geldhof, P., Peelaers, I., Claerebout, E., Berx, G., and Vercruysse, J. 2003. Identification of excretory-secretory products

of larval and adult *Ostertagia ostertagi* by immunoscreening of cDNA libraries. Molecular and Biochemical Parasitology 126:201–208.

Vinogradov, S. N., Hoogewijs, D., Bailly, X., Arredondo-Peter, R., Gough, J., Dewilde, S., Moens, L., and Vanfleteren, J. R. 2006. A phylogenomic profile of globins. BMC Evolutionary Biology 6:31–48.

Vinogradov, S. N., and Moens, L. 2008. Diversity of globin function: Enzymatic, transport, storage, and sensing. Journal of Biological Chemistry 283:8773–8777.

Wasmuth, J., Schmid, R., Hedley, A. and Blaxter, M. 2008. On the extent and origins of genic novelty in the phylum nematoda. PLoS Neglected Tropical Diseases 2:e258.

Wicks, S. R., Yeh, R. T., Gish, W. R., Waterston, R. H. and Plasterk, R. H. A. 2001. Rapid gene mapping in *Caenorhabditis elegans* using a high density polymorphism map. Nature 28:160–164

Yang, J., Kloek, A. P., Goldberg, D. E., and Mathews, F. S. 1995. The structure of *Ascaris* hemoglobin domain I at 2.2 Å resolution: Molecular features of oxygen avidity. Proceedings of the National Academy of Sciences U.S.A. 92:4224–4228.